JOURNAL OF MATHEMATICAL ANALYSIS AND APPLICATIONS 144, 450–465 (1989)

Separable Value Functions for Infinite Horizon Average Reward Markov Decision Processes

D. J. WHITE

Department of Systems Engineering, University of Virginia, Charlottesville, Virginia 22901

Submitted by E. Stanley Lee

Received May 19, 1988

1. INTRODUCTION

Mendelssohn [5, 6] examines both finite and infinite horizon discounted versions of the following problem, where we have slightly changed the notation to fit in with later developments.

A population of a species can be categorized into one of p classes, with x_i^n being the size of the population in category i at the beginning of period n from the end of the time horizon. Decisions have to be made concerning the amount, y_i^n, to be left for breeding for the next year, the remainder, $x_i^n - y_i^n$ being harvested. The population sizes at the beginning of the next period take the form

$$x_{i+1}^{n-1} = g_{i+1}(y_i^n, D_i^{n-1}), \qquad 1 \leqslant i \leqslant p-1$$
$$x_1^{n-1} = g_1(D_1^{n-1}),$$

where $D^{n-1} = (D_1^{n-1}, D_2^{n-1}, ., ., .D_p^{n-1})$ is a random variable, with realised values in a set D.

The immediate reward in the current period is a function $f(\cdot)$ of $(x-y)$, viz.

$$r(x, y) = \sum_{i=1}^{p} r_i(x_i, y_i) = \sum_{i=1}^{p} f_i(x_i - y_i).$$

For a finite horizon problem with n periods remaining, if $x = (x_1, x_2, ., ., ., x_p)$ and $v^n(x)$ is the maximal expected discounted reward over the next n periods, beginning in state x, then $\{v^n(\cdot)\}$ is the unique solution to the optimality equation

$$v^n(x) = \max_{0 \leqslant y \leqslant x} \left[\sum_{i=1}^{p} r_i(x_i, y_i) + \alpha \sum_{d \in D} q(d) v^{n-1}(g(y, d)) \right], \qquad v^0(x) = 0,$$

0022-247X/89 $3.00

where the ith component of the vector $g(y, d)$ is $g_{i+1}(y_i, d_i)$, $q(d)$ is the probability that $D^{n-1} = d$, and α is the discount factor.

The reward function is separable. At the same time the state transformation has a very special structure, viz.

$$x_i^n \xrightarrow{(y_i^n, D_i^{n-1})} x_{i+1}^{n-1}, \qquad 1 \leqslant i \leqslant p-1,$$
$$x_p^n \to \phi.$$

This transformation induces a correspondence χ between $\{1, 2, ., ., ., p\}$ and $\{2, 3, ., ., ., p, \phi\}$ by

$$\chi_i = \{i+1\}, \qquad 1 \leqslant i \leqslant p-1,$$
$$\chi_p = \phi.$$

Under these conditions Mendelssohn shows (Theorem 1, based upon Veinott [8]) that

$$v^n(x) = \sum_{i=1}^{p} v_i^n(x_i),$$

where, for each i, $\{v_i^n(\cdot)\}$ is the unique solution to the following equations for $1 \leqslant i \leqslant p-1$,

$$v_i^n(x_i) = \max_{0 \leqslant y_i \leqslant x_i} \left[r_i(x_i, y_i) + \alpha \sum_{d \in D} q(d) v_{i+1}^{n-1}(g_{i+1}(y_i, d_i)) \right],$$
$$v_i^0(\cdot) = 0,$$
$$v_p^n(\cdot) = 0.$$

Limiting results, when n tends to ∞, take a similar form.

Lovejoy [4] generalizes these results, in abstract, for the more general case when the transformation χ is one $\to$ many, rather than one $\to$ one; i.e., in terms of the component indices $\{i\}$, χ is point $\to$ set for each i. With the same separability conditions on the rewards, Lovejoy shows that the separable equations take the following form for each i, where we have replaced the expectation operator by the inclusion of actual probabilities $\{q(d)\}$,

$$v_i^n(x_i) = \max_{y_i \in \omega_i(x_i)} \left[r_i(x_i, y_i) + \alpha \sum_{d \in D} q(d) \sum_{j \in \chi_i} v_j^{n-1}(g_j(x_i, y_i, d)) \right],$$
$$1 \leqslant i \leqslant p,$$

where, more generally, the action space for x_i is $\omega_i(x_i)$.

Lovejoy uses the usual value iteration procedure to show that similar results hold for the infinite horizon case.

In this paper we will look at the average reward problem for infinite horizon, finite state, Marov decision processes. In the case of Lovejoy [4], no difficulty is encountered with infinite state sets providing the reward functions are bounded (an assumption which Lovejoy makes). This is not so with the average reward case, and we will impose the finite state space requirement on our problem. In practice, one would wish to bound the state set, in any event, as an approximation if not because physical state sets are finite. We will assume that some mechanism is introduced to retain the finiteness condition, for example, by modifying the transformation when $\|x\|$ is large enough. We will illustrate this with an elementary inventory problem where, in this case, an exact optimal solution, for any finite set of starting conditions, is obtained by such a device. The main results of this paper are as follows.

(a) (Theorem 1) If the separated optimality equations have a solution, then both the gain functions and the bias functions are separable.

(b) (Theorem 2) Under certain conditions, both the gain function and the bias function are separable, and the problem can be solved by solving the separated optimality equations.

(c) (Theorem 3) The gain function is always separable.

(d) The special structure of the χ transformation enables further decomposition of the index set $\{i\}$ into equivalence classes, which facilitates the solution procedures.

An elementary inventory control problem is considered.

We begin with the framework for our class of problems.

2. The Framework

In order to maintain notational consistency with Lovejoy [4] we will restate our separability assumption as he does, with some slight simplification since our sets will be finite, and some of Lovejoy's requirements are automatically satisfied. In the following, $\mathbf{X}$ is used to denote a vector cross product, and $\Omega(x)$ is the feasible action set for x.

(a) The state and control spaces S, C are finite and can be partitioned into p components

$$S = \bigtimes_{j=1}^{p} S_j, \qquad C = \bigtimes_{j=1}^{p} C_j$$

with $S \subseteq R^{n_j}$, $C_j \subseteq R^{m_j}$ for some integers $\{n_j, m_j\}$, $1 \leqslant j \leqslant p$, and S, C are finite.

(b) $\Omega(x) = \mathbf{X}_{j=1}^{p} \omega_j(x_j)$, $x \in S$, $x_j \in S_j$ for all j, here, for each j and $x_j \in S_j$, $\omega_j(x_j) \subseteq C_j$ is non-empty.

(c) For each $i \in \{1, 2, ., ., ., p\}$, there is a set $\chi_i \subseteq \{1, 2, ., ., ., p\}$ such that

$$\chi_i \cap \chi_j = \phi \qquad \text{if} \quad i \neq j$$

$$\bigcup_{i=1}^{p} \chi_i \subseteq \{1, 2, ., ., ., p\}.$$

(d) $r(x, y) = \sum_{j=1}^{p} r_j(x_j, y_j)$, where

$$x = \mathop{\mathbf{X}}_{j=1}^{p} x_j, \qquad y = \mathop{\mathbf{X}}_{j=1}^{p} y_j$$

and $r(\cdot, \cdot)$ is the reward function.

(e) For each $x \in S$, $y \in \Omega(x)$, $d \in D$,

$$g(x, y, d) = \mathop{\mathbf{X}}_{j=1}^{p} g_j(x_{i(j)}, y_{i(j)}, d),$$

where $i(j) = i$ for $j \in \chi_i$, $1 \leqslant i \leqslant p$, and D is finite.

In (b) we have required that $\omega_j(x_j) \neq \phi$. As we shall see in our inventory problem, for some j this requires a purely nominal action to ensure meaningful equations.

In (c) we will allow $\chi_i = \phi$ for some i, which is analogous to an absorbing state condition.

These are slight deviations from the framework of Lovejoy.

The problem is to find a policy to maximize the expected reward per period in the long run.

We now have a finite state, finite action, Markov decision process, which gives rise to a unique solution of the following equations (see Derman [2, p. 72]),

$$\begin{aligned} w(x) + \theta(x) &= \max_{y \in \Omega(x)} \left[r(x, y) + \sum_{d \in D} q(d) w(g(x, y, d)) \right] \\ \theta(x) &= \max_{y \in \Omega(x)} \left[\sum_{d \in D} q(d) \theta(g(x, y, d)) \right], \end{aligned} \tag{OE}$$

subject to $w(x) = 0$ for some x in each of the ergodic classes generated by an optimal decision rule. $\theta(\cdot)$ is the gain function. With the condition

$w(x) = 0$ for some x, $w(\cdot)$ is the bias function with an unknown constant added. We will, however, refer to $w(\cdot)$ as the bias function, since the unknown constant does not change the relative values of $\{w(x)\}$, $x \in S$.

3. Separability of the Gain and Bias Functions

We may now use two approaches to the separability issue.

The first merely requires that the appropriate equations have a solution. This can be checked by trying to solve the equations.

The second gives conditions under which the separability results must hold. These conditions are, however, in general difficult to verify. We present our results as three theorems. First we quote the separable optimality equations, $1 \leqslant i \leqslant p$,

$$w_i(x_i) + \theta_i(x_i) = \max_{y_i \in \omega_i(x_i)} \left[r_i(x_i, y_i) + \sum_{j \in \chi_i} \sum_{d \in D} q(d) w_j(g_j(x_i, y_i, d)) \right],$$
$$\theta_i(x_i) = \max_{y_i \in \omega_i(x_i)} \left[\sum_{j \in \chi_i} \sum_{d \in D} q(d) \theta_j(g_j(x_i, y_i, d)) \right], \qquad \text{(OE)(i)}$$

with $\sum_{i=1}^{p} w_i(x_i) = 0$ for some x in each ergodic class.

Theorem 1. *If the equations* $\{\text{(OE)(i)}\}$ *have a solution, for* $1 \leqslant i \leqslant p$, *then*

$$w(\cdot) = \sum_{i=1}^{p} w_i(\cdot),$$
$$\theta(\cdot) = \sum_{i=1}^{p} \theta_i(\cdot). \qquad \text{(SE)}$$

Proof. Equation (OE) has a unique solution, subject to the stipulated ergodic class requirements. Hence, if we substitute $(w(\cdot), \theta(\cdot))$, given by (SE), into (OE), and if we can solve the resulting equations, this will be the required solution.

We begin with the $\theta(\cdot)$ equations. The right-hand side of (OE) takes the following form:

$$\max_{\substack{\{y_i \in \omega_i(x_i)\} \\ 1 \leqslant i \leqslant p}} \left[\sum_{d \in D} q(d) \sum_{j=1}^{p} \theta_j(g_j(x_{i(j)}, y_{i(j)}, d)) \right]$$
$$= \max_{\substack{\{y_i \in \omega_i(x_i)\} \\ 1 \leqslant i \leqslant p}} \left[\sum_{i=1}^{p} \sum_{j \in \chi_i} \sum_{d \in D} q(d) \theta_j(g_j(x_i, y_i, d)) \right]$$
$$= \sum_{i=1}^{p} \max_{y_i \in \omega_i(x_i)} \left[\sum_{j \in \chi_i} \sum_{d \in D} q(d) \theta_j(g_j(x_i, y_i, d)) \right].$$

Separating out the separate i-components, and equating with the i-component on the left-hand side, gives the requisite (OE)(i) $\theta(\cdot)$ equation. The (OE)(i) $(w(\cdot), \theta(\cdot))$ equation is likewise derived, the extra component in the proof requiring the use of the separability of $r(\cdot, \cdot)$ (see (d)). ∎

Let us now consider the second approach. In order to do this we need to introduce the finite horizon scheme.

Let

$$v_0(\cdot): S \to \mathscr{R} \text{ be arbitrary.}$$

Define the sequence $\{v_n(\cdot)\}: S \to \mathscr{R}$ as follows for $n \geqslant 1$:

$$v_n(\cdot) = Tv_{n-1}(\cdot),$$

where, for any $u(\cdot): S \to \mathscr{R}$,

$$[Tu](x) = \max_{y \in \Omega(x)} \left[r(x, y) + \sum_{d \in D} q(d) u(g(x, y, d)) \right].$$

Schweitzer and Federgruen [7, Theorem 5.5] discuss the asymptotic behavior of the sequence $\{v_n(\cdot)\}$. They list seven conditions (which we will henceforth refer to as the S.F.-Conditions) sufficient for the following to be true:

$$v_n(\cdot) - n\theta(\cdot) - w(\cdot) \text{ tends to zero as } n \text{ tends to infinity.} \qquad \text{(A)}$$

$\theta(\cdot)$ is the optimal gain function and $w(\cdot)$ is the bias function.

We may now use the asmptotic result (A) to prove the following theorem.

THEOREM 2. *If any of the S.F.-Conditions hold, and if $v_0(\cdot)$ is separable, then the separability equations* (SE) *also hold.*

Proof. The separability of $\{v_n(\cdot)\}$ is easily established inductively, given that $v_0(\cdot)$ is separable (this is the mode of proof for Lovejoy for the discounted case).

From the asymptotic result (A) we then have the following:

$$\theta(x) = \operatorname*{Lim}_{n \to \infty} [v_n(x)/n] = \operatorname*{Lim}_{n \to \infty} \left[\sum_{i=1}^{p} v_{ni}(x_i)/n \right].$$

We cannot automatically interchange the limits and the summation on the right-hand side, since it is possible that $\operatorname{Lim}_{n \to \infty} [v_{ni}(x_i)/n]$ might not exist for some i, and some $x_i \in S_i$.

Let us fix $x^0 \in S$ as a datum point.

Let

$$v_{ni}(x_i)/n = \theta_{ni}(x_i),$$

$$\theta_n(x) = \sum_{i=1}^{p} \theta_{ni}(x_i).$$

Then

$$\theta_{n1}(x_1) - \theta_{n1}(x_1^0) = \theta_n(x_1, x_2^0, ., ., ., x_p^0) - \theta_n(x_1^0, x_2^0, ., ., ., x_p^0).$$

Then, using a similar analysis for each i, we have the following:

$$\operatorname*{Lim}_{n \to \infty} [\theta_{ni}(x_i) - \theta_{ni}(x_i^0)] = \Delta_i(x_i) \text{ exists for all } i.$$

Then

$$\operatorname*{Lim}_{n \to \infty} \left[\sum_{i=1}^{p} \theta_{ni}(x_i) \right] - \operatorname*{Lim}_{n \to \infty} \left[\sum_{i=1}^{p} \theta_{ni}(x_i^0) \right] = \sum_{i=1}^{p} \Delta_i(x_i).$$

Hence the following holds:

$$\theta(x) = \theta(x^0) + \sum_{i=1}^{p} \Delta_i(x_i)$$

$$= \sum_{i=1}^{p} (\Delta_i(x_i) + \theta(x^0)/p)$$

$$= \sum_{i=1}^{p} \theta_i(x_i), \text{ say.}$$

Thus $\theta(\cdot)$ satisfies the separability requirement.

Using the asymptotic result (A), a similar analysis shows that $w(\cdot)$ also satisfies the separability requirement. ∎

The S.F.-Conditions may be difficult to verify in general. We do have a weaker theorem, which establishes the general separability of $\theta(\cdot)$, but not of $w(\cdot)$.

This result derives from a weaker form of the asymptotic result (A). Brown [1, Theorem 4.3] shows the following asymptotic result to hold for all finite action, finite state, problems:

$$v_n(\cdot) - n\theta(\cdot) \text{ is bounded.} \qquad \text{(B)}$$

Using this result, and a similar analysis to that of Theorem 2, we automatically have the following theorem, for which no proof will be given.

THEOREM 3. *The $\theta(\cdot)$ part of the separability equations* (SE) *holds.*

This result also comes automatically from Lovejoy's result by noting that, if $v_\alpha(\cdot)$ is the value function in the discounted case, then

$$\theta(\cdot) = \lim_{\alpha \to 1-} [(1-\alpha)\, v_\alpha(\cdot)] .$$

Even to obtain this result, the separability of a limiting sequence of separable functions still has to be extablished as in Theorem 2.

4. COMPUTATIONAL ASPECTS

Let us now suppose that we wish to solve the optimality equations $\{\mathrm{OE(i)}\}$, assuming that they have a solution with $\sum_{i=1}^{p} w_i(x_i) = 0$ for some x in each ergodic class.

Then these equations may be solved, in principle, by the policy space method of Howard [3, p. 64].

However, the particular structure of the problem offers some simplification, which we may, in some cases, exploit to improve the computational efficiency of the algorithms used.

The $\{\chi_j\}$ sets may be used to generate directed graphs rooted at any node, where, if G_i is such a graph, rooted at node i, $\mathrm{arc}(j, k) \in G_i$ if and only if $k \in \chi_j$, and $j \in G_i$.

Let us consider any path from node i in G_i. Since $\chi_j \cap \chi_k = \phi$ if $j \neq k$, the only node which can occur twice on this path is node i, and at most one such path can have this property. Since S is finite, all other paths culminate in a node j where $\chi_j = \phi$.

Thus we can classfy nodes into two categories, viz.

S^1: the set of nodes for which the subgraphs generated by the χ transformation contain a path beginning and ending at that node;

S^2: the set of nodes for which the subgraphs generated by the χ transformation terminate at nodes $j \in S$ for which $\chi_j = \phi$.

Since the objective is to find policies to maximize the $\{\theta(x)\}$ values, the nodes in the set S^2 may be ignored.

The nodes in the set S^1 may be partitioned into equivalence classes as follows.

Let $i \in S^1$, and let G_i be the associated subgraph generated by the transformation χ. Let P_i be the unique path in G_i leading from the root i back to i (the path is a simple cycle). Then we say

$$i \sim j \text{ if and only if } j \in P_i.$$

It is early seen that $i \sim j$, if and only if $j \sim i$ and, if $i \sim j$, $j \sim k$, then $i \sim k$.

In effect, each node in an equivalence class generates the same cyclic path from itself back to itself.

Let $\mathscr{P}$ be the set of cyclic paths determined by the equivalence classes.

Then we can solve our problem by solving a set of independent problems, each relating to one equivalence class.

As an example, suppose we have the following:

$$S = \{1, 2, 3, ., ., ., 9\},$$

$$\chi_1 = \{2\}, \quad \chi_2 = \{3\}, \quad \chi_3 = \{1, 4, 5\}, \quad \chi_4 = \phi, \quad \chi_5 = \phi,$$

$$\chi_6 = \{7\}, \quad \chi_7 = \{\phi\}, \quad \chi_8 = \{6, 9\}, \quad \chi_9 = \{8\}.$$

Then the two equivalence classes in S^1 are

$$\{1, 2, 3\}, \quad \{8, 9\}.$$

The paths corresponding to these are as follows: $\mathscr{P} = \{1 \to 2 \to 3 \to 1), (8 \to 9 \to 8)\}$.

The optimality equations may now be simplified as follows, noting that we now use $\tilde{w}$ instead of w, since the bias terms may well be different when we discard parts of the set S.

Let $i \in S^1$ and let $j(i)$ be the successor node of i in the path P_i. For all nodes i in the same equivalence class we have the following optimality equations.:

$$\tilde{w}_i(x_i) + \theta_i(x_i) = \max_{y_i \in \omega_i(x_i)} \left[r_i(x_i, y_i) + \sum_{d \in D} q(d)\, \tilde{w}_{j(i)}(g_{j(i)}(x_i, y_i, d)) \right]$$

$$\theta_i(x_i) = \max_{y_i \in \omega_i(x_i)} \left[\sum_{d \in D} q(d) \theta_{j(i)}(g_{j(i)}(x_i, y_i, d)) \right]. \qquad \widetilde{\mathrm{OE}}(\mathrm{i})$$

We also require $\sum_{j \in P_i} w_j(x_j) = 0$ for some x in each ergodic class.

Once the optimality equations $\{\widetilde{\mathrm{OE}}(\mathrm{i})\}$ have been solved for $\{\theta_i(\cdot)\}$, these may be substituted into the optimality equations $\{\widetilde{\mathrm{OE}}(\mathrm{i})\}$ if it is further required to solve for the bias functions $\{w_i(\cdot)\}$. It is to be noted that, for $i \in S^2$, the $\{w_i(\cdot)\}$ functions can be calculated by working backwards from the terminal node. If $i \in S^2$ and if i is a terminal node (so $\chi_i = \phi$), then

$$w_i(x_i) = \max_{y_i \in \omega_i(x_i)} [r_i(x_i, y_i)].$$

At the beginning of this section we assumed that we wished to solve the optimality equations $\{\widetilde{\mathrm{OE}}(\mathrm{i})\}$. However, if the problem is essentially to find

a policy to optimize the gain functions $\{\theta_i(\cdot)\}$ we may use Theorem 3, and then it is sufficient to solve the equations $\{\widetilde{\mathrm{OE}}(\mathrm{i})\}$.

In this paper we will not deal with the general solution to $\{\widetilde{\mathrm{OE}}(\mathrm{i})\}$. Let us, however, look at the case when, for a given equivalence class of nodes in S^1, we have only a single ergodic chain for all the policies. In this case $\theta_i(x_i)=\theta$ for all nodes i in the equivalence class, and for all $x_i \in S_i$.

Let us number the nodes in a specific cyclic path as $i=1, 2, ., ., ., m$, with $i+1$ being the successor of i, and identifying $m+1$ with 1.

Let us now define the functions and vectors

$$\tilde{y}_t(\cdot)=(y_t(\cdot), y_2(\cdot), ., ., ., y_m(\cdot)), \qquad 1 \leqslant t \leqslant m,$$

$$\tilde{d}_t=(d_1, d_2, ., ., ., d_t), \qquad 1 \leqslant t \leqslant m,$$

$$\tilde{g}_t(\cdot, \tilde{y}_{t-1}(\cdot), \tilde{d}_{t-1})=g_t(\tilde{g}_{t-1}(\cdot, \tilde{y}_{t-2}(\cdot), \tilde{d}_{t-2}), y_{t-1}(\cdot), d_{t-1}),$$

$$2 \leqslant t \leqslant m+1 \equiv 1$$

$$\tilde{g}_1(x_1, \tilde{y}_0(\cdot), d_0) \equiv x_1,$$

where

$$y_t(\cdot): S_t \to C_t,$$

$$\tilde{g}_t(\cdot, \tilde{y}_{t-1}(\cdot), \tilde{d}_{t-1}): S_1 \times \prod_{u=1}^{t-1} C_u \times D^{t-1} \to S_t.$$

The function $\tilde{g}_t(., ., .,)$ in effect, determines x_t when the decision rules up to node $t-1$ (inclusive) and the random variables up to node t (inclusive) and x_1 are given.

We also define the m step reward function as

$$\tilde{r}(\cdot, \tilde{y}_m(\cdot))=\sum_{t=1}^{m} \sum_{\tilde{d}_m \in D^m} \tilde{q}(\tilde{d}_m) \tilde{r}_t(\tilde{g}_t(\cdot, \tilde{y}_{t-1}(\cdot), \tilde{d}_{t-1}), y_t(\cdot)),$$

where $\tilde{q}(\tilde{d}_m)=\prod_{t=1}^{m} q(d_t)$.

Then, for a given policy, $\tilde{y}_m(\cdot)$, the equations for determining the gain θ are as follows:

$$\tilde{w}_1(\cdot)+m\theta=\tilde{r}(\cdot, \tilde{y}_m(\cdot))+\sum_{\tilde{d}_m \in D^m} \tilde{q}(\tilde{d}_m) \tilde{w}_1(\tilde{g}_1(\cdot, \tilde{y}_m(\cdot), \tilde{d}_m)). \quad (\widetilde{\mathrm{OE}})(1)$$

To solve $((\widetilde{\mathrm{OE}})(1)$ by the policiy space method of Howard [3] we proceed as follows.

(i) Select a policy $\tilde{y}_m^1(\cdot)$.

(ii) Solve the equations $(\widetilde{\mathrm{OE}})(1)$ for $(\tilde{w}_1^1(\cdot), \theta^1)$, setting $\tilde{w}_1(x_1)=0$ for some $x_1 \in S_1$.

(iii) Find $\tilde{y}_m^2(\cdot)$ to maximize the right-hand side of $(\widetilde{OE})(1)$ when $\tilde{w}_1(\cdot)$ is replaced by $\tilde{w}_1^1(\cdot)$.

(iv) Replace $\tilde{y}_m^1(\cdot)$ in (i) by $\tilde{y}_m^2(\cdot)$ and continue until the sequence $\{\tilde{w}_1^n(\cdot)\}$ has converged, which it will do in a finite number of steps.

In step (iii), for each $x_1 \in S_1$, the step is equivalent to selecting a super-action $(y_1, y_2(\cdot), y_3(\cdot), ., ., ., y_m(\cdot)))$ from among a finite set. However, for the physical problem on hand, we require that $y_t(\cdot)$ be independent of $\{y_1, y_2(\cdot), y_3(\cdot), ., ., ., y_{t-1}(\cdot)\}$. Such a solution exists and is found by the following scheme.

Define the sequence $\{\tilde{\tilde{w}}_t(\cdot)\}$ as

$$\tilde{\tilde{w}}_t(x_t) = \max_{y_t \in \omega_t(x_t)} \left[r_t(x_t, y_t) + \sum_{d \in D} q(d) \tilde{\tilde{w}}_{t+1}(g_{t+1}(x_t, y_t, d)) \right],$$

$$1 \leqslant t \leqslant m$$

with

$$\tilde{\tilde{w}}_{m+1}(\cdot) \equiv \tilde{w}_1^1(\cdot).$$

It is readily seen that this scheme will execute step (iii) and, at the same time, maintain the independence property required.

5. An Inventory Problem

Consider the following inventory situation.

(i) A single commodity has to be purchased at the beginning of each unit of time.

(ii) There is a lead time of p time units for the delivery of any amount purchased.

(iii) If x is the inventory level at the beginning of any time unit, inclusive of recent delivery, there is an expected cost for the time unit of $c(x)$, where the cost $c(x)$ includes inventory holding costs and backlog costs. x may be negative, representing an existing backlog situation at the beginning of the time unit.

(iv) The cost of purchasing a quantity z is $a(z)$.

(v) The demand in any unit time interval is discrete and is independently and identically distributed in each time unit, with probability $q(d)$ that the demand is equal to d, $0 \leqslant d \leqslant \bar{d} < \infty$.

(vi) The problem is to find a purchasing policy to minimize the expected cost per unit time over an infinite time horizon.

Note that our general framework is in maximization form to conform with Lovejoy [4]. However, minimization problems clearly also fit into the framework.

In order to keep the problem simple and to keep it within the bounds of finite state space analysis, we need to impose some further conditions as mentioned in Section 1. In order to do this we need first of all to define our state space and action space.

Let:

x_i be the inventory on hand plus the inventory on order up to and including the start of the ith time unit from the current decision epoch, excluding any order made at that decision epoch, $1 \leqslant i \leqslant p$;

y ($\geqslant x_p$) be the new level to which x_p will be increased (so that the order quantity is $y - x_p$ at the current decision epoch).

We now add the following condition.

(vii) There is a $\bar{z} < \infty$, and a pair $\{\underline{x}, \bar{x}\}$, with $-\infty < \underline{x} < \bar{x} < \infty$, such that $0 \leqslant z \leqslant \bar{z}$ and $\underline{x} \leqslant x_i \leqslant \bar{x}$, $1 \leqslant i \leqslant p$.

For any given initial state x, it is clearly possible to choose $\{\bar{z}, \underline{x}, \bar{x}\}$, and subsequent actions, in order to ensure that all future states lie within the specified bounds. However, our state space S is a complete product space, and we have to impose some structure on the problem in order to keep it within bounds. We will proceed as follows.

We first of all note the structure. We have the following:

$$\chi_1 = \phi, \qquad \chi_i = \{i-1\}, \quad 2 \leqslant i \leqslant p-1, \qquad \chi_p = \{p-1, p\}.$$

If x' is the transform of x, for a given demand level d, we have the following:

$$x'_i = x_{i+1} - d, \qquad 1 \leqslant i \leqslant p-1,$$
$$x'_p = y - d.$$

In order to keep within the finite state space structure specified by the bounds $\{\bar{z}, \underline{x}, \bar{x}\}$, we modify the transformations as follows, where $\{x''_i\}$ are the modified transformed states.

For $1 \leqslant i \leqslant p-1$,

$$x''_i = \max[\underline{x}, x_{i+1} - d] = g_i(x_{i+1}, d).$$

In addition to this, we also impose an extra constraint on y as follows:

(viii) $y \leqslant \bar{x}$ and if $\underline{x} \leqslant x_p \leqslant \underline{x} + \bar{d}$, then $y \geqslant x_p + \bar{d}$.

We let $\omega_p(x_p)$ be the admissible region for y, and

$$g_p(x_p, y, d) = y - d.$$

With these modifications, our framework will be within the finite state, finite action requirements, and, for any specified initial state, the bounds may be chosen so that feasible paths exist. In addition, if the bounds are large enough one would expect the modifications not to unduly restrict the decision rule space. Once a solution has been obtained it is possible to check the assumptions and determine whether or not any loss of optimality has risen. In any event, even for an infinite state problem, finite approximation models are required to solve it.

With this model, the average cost optimality equations, before separation, are as follows, where $w(\cdot)$, $\theta(\cdot)$ are, respectively, the bias function and the average cost function:

$$w(x) + \theta(x) = \min_{y \in \omega_p(x_p)} \left[c(x_i) + a(y - x_p) + \sum_{d=0}^{d} q(d) w(g(x, y, d)) \right],$$

where $g(.,.,.)$ has been specified earlier on. The separation equations are then as follows where (see Section 4) $S^1 = \{p\}$, $S^2 = \{1, 2, ., ., ., p-1\}$:

$$w_1(x_1) + \theta_1(x_1) = c(x_1),$$

$$w_i(x_i) + \theta_i(x_i) = \sum_{d=0}^{d} q(d) w_{i-1}(g_{i-1}(x_i, d)), \qquad 2 \leqslant i \leqslant p-1,$$

$$w_p(x_p) + \theta_p(x_p) = \min_{y \in \omega_p(x_p)} \left[a(y - x_p) + \sum_{d=0}^{d} q(d) w_{p-1}(g_{p-1}(x_p, d)) + \sum_{d=0}^{d} q(d) w_p(y - d) \right],$$

$$\theta_1(x_i) = 0,$$

$$\theta_i(x_i) = \sum_{d=0}^{d} \theta_{i-1}(g_{i-1}(x_i, d)), \qquad 2 \leqslant i \leqslant p-1,$$

$$\theta_p(x_p) = \min_{y \in \omega_i(x_i)} \left[\sum_{d=0}^{d} q(d) \theta_{p-1}(g_{p-1}(x_p, d)) + \sum_{d=0}^{d} q(d) \theta_p(y - d) \right].$$

From these we see that $\theta_i = 0$, $1 \leqslant i \leqslant p-1$. The equations then reduce

to the following form, in accordance with the decomposition given in Section 4:

$$w_1(x_1) = c(x_1),$$

$$w_i(x_i) = \sum_{d=0}^{d} q(d) w_{i-1}(q_{i-1}(x_i, d)), \qquad 2 \leqslant i \leqslant p-1,$$

$$w_p(x_p) + \theta_p(x_p) = \min_{y \in \omega_p(x_p)} \left[a(y - x_p) + \sum_{d=0}^{d} w_{p-1}(g_{p-1}(x_p, d)) + \sum_{d=0}^{d} q(d) w_p(y-d) \right],$$

$$\theta_p(x_p) = \min_{y \in \omega_p(x_p)} \left[\sum_{d=0}^{d} q(d) \theta_p(y-d) \right].$$

In addition, we must set $\sum_{i=1}^{p} w_i(x_i) = 0$ for some value of x in each of the ergodic classes of the Markov processes defined by the policies.

It is to be noted that, for all realisable states, we have $x_1 \leqslant x_2 \leqslant ., ., ., ., \leqslant x_p$. The solution to the above set of equations will, in general, involve non-realisable states. However, this will not matter.

Let us now assume, without loss, that $\underline{x} \leqslant \bar{x} + \bar{d}$. Define the sequence of function $h_{p-k}(\cdot)$: $\mathscr{R}^{k+1} \to \mathscr{R}$ as follows:

$$h_{p-k}(x_p, s_1, s_2, ., ., ., s_k) = g_{p-k}(h_{p-k+1}(x_p, s_1, s_2, ., ., ., s_{k-1}), s_k), \qquad 1 \leqslant k \leqslant p-1,$$

$$h_p(x_p) \equiv x_p.$$

Now let $\underline{x} + (p-1)\bar{d} \leqslant x_p \leqslant \bar{x}$.

It is then easily seen, inductively, that the following holds:

$$\underline{x} \leqslant h_{p-k} \leqslant \bar{x}, \qquad 0 \leqslant k \leqslant p.$$

As a consequence of this, for $x_p \in [\underline{x} + (p-1)\bar{d}, \bar{x}]$ we can simplify the problem. The following holds:

$$\sum_{d=0}^{d} q(d) w_{p-1}(g_{p-1}(x_p, d)) = \sum_{d=0}^{(p-1)d} Q_{p-1}(d) c(x_p - d),$$

where $Q_{p-1}(d)$ is the probability that the total demand over $p-1$ unit time intervals is equal to d.

We may now state the result in the following form.

THEOREM 4. *For $x_p \in [\underline{x} + (p-1)\bar{d}, \bar{x}]$, the $(w(\cdot), \theta(\cdot))$ solution for the x_p-state component vector is the unique solution to the following equations:*

$$w_p(x_p) + \theta_p(x_p) = \min_{y \in \omega_p(x_p)} \left[a(y - x_p) + \sum_{d=0}^{(p-1)\bar{d}} Q_{p-1}(d) c(x_p - d) + \sum_{d=0}^{\bar{d}} q(d) w_p(y-d) \right],$$

$$\theta_p(x_p) = \min_{y \in \omega_p(x_p)} \left[\sum_{d=0}^{\bar{d}} q(d) \theta_p(y-d) \right],$$

with $w_p(x_p) = 0$ for some x_p in each of the ergodic classes.

6. SUMMARY AND COMMENTS

The purpose of this paper has been to extend the results of Lovejoy [4] from infinite horizon discounted problems to infinite horizon average reward problems.

Theorem 1 shows that if the separated optimality equations have a solution, the problem will have a separable solution for $(w(\cdot), \theta(\cdot))$. Theorem 2 shows that, under certain conditions, the solution to our problem must be separable for $(w(\cdot), \theta(\cdot))$. Theorem 3 shows that the gain function $\theta(\cdot)$ is always separable. This is all done in Section 3.

Section 4 shows how the particular form of the transformation function χ: $S \to 2^S$ may be used to aid the solution to the problem by decomposing S into equivalence classes and, first of all, finding a solution to each cyclic path problem corresponding to each equivalence class.

Section 5 treats an elementary inventory problem within this framework, although the main optimality equation in x_p may be derived from first principles.

There are some outstanding questions, viz.

(i) Under what conditions, more general than the S.F.-conditions given in Theorem 2, do the separated optimality equations have a solution?

This question is largely of academic interest, since, if the objective is to find any gain optimal solution, we may use Theorem 3 and the computational procedures of Section 4 using artificial bias functions $\{\tilde{w}_i(\cdot)\}$.

(ii) What computational routines might be developed, using the equivalence class structure of Section 4, to handle problems with multiple ergodic classes?

A crucial property is that of $\chi_i \cap \chi_j = \phi$ if $i \neq j$. There are some problems

which almost, but not quite, satisfy these requirements. This leads us to the next question.

(iii) Is it possible to develop some special schemes for problems for which $\chi_i \cap \chi_j \neq \phi$ for some pairs (i, j)?

Finally, it is to be noted that the equivalence class ideas and associated computational schemes clearly apply, in a suitably modified form, to the discounted problems, and to finite horizon problems, with or without discount factors.

REFERENCES

1. B. BROWN, On the iterative method of dynamic programming on a finite state space discrete time Markov process, *Ann. Math. Statist.* **36** (1965), 1279–1285.
2. C. DERMAN, "Finite State Markovian Decision Processes," Academic Press, San Diego, 1970.
3. R. A. HOWARD, "Dynamic Programming and Markov Processes," Wiley, New York, 1960.
4. W. S. LOVEJOY, Policy bounds for Markov decision processes, *Oper. Res.* **34** (1986), 630–637.
5. R. MENDELSSOHN, Optimal Harvesting strategies for stochastic single species multistage class models, *Math. Biosci.* **41** (1978), 159–174.
6. R. MENDELSSOHN, Managing stochastic multispecies models, *Math. Biosci.* **49** (1980), 249–261.
7. P. J. SCHWEITZER AND A. FEDERGRUEN, The asymptotic behavior of undiscounted value iteration in Markov decision problems, *Math. Oper. Res.* **2** (1977), 360–381.
8. A. F. VEINOTT, Optimal policy for a multi-product, dynamic non-stationary, inventory problem, *Management Sci.* **12** (1965), 206–222.